連濁の生起率に基づく日本語複合語の分類 : 連濁データベースによる研究

著者	太田 聡, 太田 真理
雑誌名	国立国語研究所論集
号	10
ページ	179-191
発行年	2016-01
URL	http://doi.org/10.15084/00000814

国立国語研究所論集 (NINJAL Research Papers) 10: 179–191 (2016)
ISSN: 2186-134X print/2186-1358 online

連濁の生起率に基づく日本語複合語の分類
——連濁データベースによる研究——

太田　聡[a]　　太田真理[b]

[a]山口大学／国立国語研究所 共同研究員
[b]東京大学

要旨

連濁はもっとも広く知られた日本語の音韻現象の１つである。先行研究では，日本語の複合語は連濁の生起率の違いに基づいて，いくつかのグループに分類されることが提案されている。しかしながら先行研究では，連濁生起率の分類基準が恣意的であった点，またグループの数をあらかじめ仮定していた点に問題があった。そこで本研究では，混合正規分布モデルに基づくクラスター分析と連濁データベース（Irwin and Miyashita 2015）を用いて，日本語複合語を分類する際の最適な分類基準とクラスター数を検討した。複合名詞と複合動詞のどちらも，２つのクラスターを仮定したモデルが最適であり，クラスター同士の分類基準は，複合名詞では連濁生起率が 90%，複合動詞では 40% であった。これらの結果は先行研究のクラスター数や分類基準とは異なるものであった。我々の結果は，モデルに基づくクラスター分析が言語データに対する最適な分類を行う上で非常に有効であることを示すものである *。

キーワード：連濁，複合語，生起率，クラスター分析，混合正規分布モデル

1. はじめに

日本語では，複合語の後部要素が有声性に関して対立を持つ無声阻害音（清音：/k/, /s/, /t/, /h/）で始まる場合に，対応する有声阻害音（濁音：/g/, /z/, /d/, /b/）に変化する連濁という現象が知られている。例えば（1a）の後部要素は清音 /h/ で始まるため, 対応する濁音 /b/ に変化する。その一方で，（1a）と（1b）の対比から明らかなように，連濁は常に生じるわけではない。

（1）　a.　ごみ ＋ はこ → ごみばこ（以下の例では，連濁を下線で示す）
　　b.　ごみ ＋ かご → ごみかご／* ごみがご

上記のような連濁の生起・非生起を説明する法則として，「複合語の後部要素が濁音を含む場合は連濁が生じない」という「本居・ライマンの法則」が提案されている（Lyman 1894）。（1b）は「かご」が濁音 /g/ を含むため，本居・ライマンの法則によって連濁が生じないことが正しく予測される。なお本研究では, 連濁が生じた複合語を「連濁形」, 連濁が生じていない複合語を「非連濁形」と呼ぶ。

* 本稿は国立国語研究所基幹型共同研究プロジェクト「日本語レキシコン―連濁事典の編纂」（プロジェクトリーダー：ティモシー J. バンス）の研究成果である。また，本稿の内容は，3rd International Conference on Phonetics and Phonology（於・国立国語研究所：2013 年 12 月）における発表 “*Rendaku* ‘enthusiasts’ and *rendaku* ‘indifferents’: Classification of compound nouns based on the frequency of *rendaku*” に基づいたものである。発表時に有益なコメントを下さった方々に感謝を申し上げる。

連濁に影響する意味的要因として，(2a) と (2b) の対比が示すように前部要素が後部要素を修飾する場合は連濁するが，両者が意味的に並列される複合語（並列複合語）では連濁しないことが知られている（中川 1966）。

(2) a. やま + かわ → やまがわ（山の川）
b. やま + かわ → やまかわ（山と川）

また，(3) のように同一の単語の繰り返しからなる複合語（畳語）では連濁が生じやすく（Lyman 1894），(4) のように，人名や地名などの固有名詞では連濁の生起に関して曖昧性が生じやすいことも知られている。

(3) ひと + ひと → ひとびと（人々）
(4) やま + さき → やまざき／やまさき（山崎：人名・地名）

さらに，「すれる」と「ずれる」のように，語頭が濁音で始まる語（濁音独自単語）が連濁形とは別に存在する場合に，両者の曖昧性を避けるために連濁が阻害される可能性も指摘されている（Irwin 2014）。語頭に濁音が許されない和語では，語頭を濁音に変化させる連濁によって，修飾部と主要部の境界を標示して主要部を際立たせる機能があると考えられる（田中 2009: 100–102）。また，語頭に濁音が許される漢語や外来語では，語頭の有声化では修飾部と主要部の境界を標示できないため連濁が許されないと考えられる。濁音独自単語も同様に，有声の語頭が連濁の結果か元から有声であったのかが曖昧なため，単語の境界を標示することができず，従って連濁が許されないと考えられる[1]。以上のような音韻的･意味的要因に基づいて連濁の生起が説明可能かどうかを，統計的多変量解析を用いて検討した我々の先行研究では，90% 程度の複合語において正しく連濁の生起が説明できることが明らかとなっている（太田 2015）。さらに，統語的要因として，(5) の対比から示されるように，「複合語の木構造中で右枝に来る要素のみで連濁が生じる」という，「右枝条件」が知られている（Otsu 1980）[2]。

(5) a. ぬりばし + いれ → [[ぬりばし][いれ]]（塗り箸専用の入れ物）
b. ぬり + はしいれ → [[ぬり][はしいれ]]（漆塗りの箸入れ）

ここまで先行研究で提案された連濁の生起に関わる代表的な要因について概観したが，これらの要因では連濁の生起が説明できない場合も存在する。

[1] なお濁音独自単語と連濁形の曖昧性が原因と考えられる現象に「世間ずれ」の意味の変遷があげられる。「国語に関する世論調査」（文化庁編 2014）によれば，本来の「世間＋すれ」（世間にあって苦労し，悪賢くなっている）という意味ではなく，「世間＋ずれ」（世の中の考えから外れている）という意味で使用される割合が，50 代以下では 50% を超えている。
[2] なお，本研究で用いた 28,800 語の複合語はいずれも後部要素自体が複合語ではないため，「右枝条件」の対象とはならない。

(6)　a.　よ + きり → よぎり（夜霧）
　　b.　よ + つゆ → よつゆ／* よづゆ（夜露）

(6) のように，連濁の生起が予測できない複合語に対して，先行研究の多くは個々の後部要素の語彙的な特性に原因を帰属させてきた。このような後部要素の違いに起因する連濁生起率の違いに基づいて，Rosen（2001）は “*rendaku* lover” と “*rendaku* hater” という 2 つのサブグループを提案している。*Rendaku* lover は連濁生起率が 66% を超える後部要素を指し，一方で，*rendaku* hater は連濁生起率が 33% を下回る後部要素を指す。また，Rosen（2001）は，どちらにも当てはまらない後部要素が非常に少ないことも指摘している。これに対して Irwin（2012）は，“*rendaku* waverer” という *rendaku* lover にも *rendaku* hater にも当てはまらない（つまり連濁生起率が 33% と 66% の間にある）後部要素が存在し，10% 程度の後部要素は *rendaku* waverer に相当することを報告している（Irwin 2012）。

これら 2 つの先行研究は，日本語の複合語の後部要素は，連濁生起率に基づいてサブグループに分類できると提案した点で画期的である。しかしながら，グループ同士を区別する連濁生起率の基準に 33% と 66% という恣意的な値を用いた点，またサブグループの数が 2 種類または 3 種類であるという暗黙の前提を置いている点で問題があった。そこで本研究では，データに基づいてサブグループに分類する統計的手法であるクラスター分析を用いて，後部要素をいくつのサブグループに分類した場合にデータとの適合度が最大になるのか，またその際に連濁生起率の分類基準はどうなるのかを検討した。なお，Rosen（2001）と Irwin（2012）では，後部要素が名詞の複合語のみを対象にしているが，我々は後部要素が動詞に由来する名詞（例：書き）や動詞の場合についても検討を行った。本研究では，連濁データベース v2.5（Irwin and Miyashita 2015）に含まれる複合語のうち，音韻的要因，統語的要因，意味的要因では連濁生起の揺れが説明できない 28,800 語を対象にした。

2. 研究方法

2.1 連濁データベース

連濁データベースは，国立国語研究所共同研究プロジェクト「日本語レキシコン―連濁事典の編纂」の一環として構築が進められており，本研究では v2.5 を使用した（Irwin and Miyashita 2015）。連濁データベースには，『広辞苑』第 6 版または『新和英大辞典』第 5 版に掲載されている見出し語のうち，以下のいずれかに該当する 34,432 個の複合語が収録されている。

(7)　後部要素が濁音を含まない和語からなる複合語（例：飲み薬）
(8)　後部要素が漢語または外来語で，連濁が生じる複合語（例：株式会社，雨ガッパ）
(9)　後部要素に濁音を含む和語で，連濁が生じる複合語（例：縄梯子）

また連濁データベースには，各複合語の連濁の有無に加え，前部要素と後部要素の語種・品詞，前部要素と後部要素の関係，人名・氏名のみで使用されるか，本居・ライマンの法則に反するか，

畳語か，並列複合語か，接尾辞か（例：～様），濁音独自単語があるか，/b/ に由来する /m/ を含むか（例：蝙蝠），という情報も記載されている。

本研究では，畳語と並列複合語については連濁の生起が意味的な要因により説明可能であると考えられるため，解析対象から除外した。また，固有名詞も連濁の生起に揺れがあると考えられるため解析対象から除外した。さらに，頻度が低い後部要素では安定して連濁の生起率を求めることが難しいため，連濁データベースへの収録数が 10 個未満の後部要素については解析の対象から除外した。以上の結果，28,800 語の複合語を解析の対象とした。このうち後部要素の異なり語数は 629 語であった（表 1）。

表 1　後部要素の品詞と連濁生起率

	後部要素の品詞		
	名詞	動詞由来名詞	動詞
異なり語数	404	168	57
合計	19,007	7,400	2,393
連濁生起率（%）	75.8	76.3	15.0

2.2 連濁生起率

連濁データベースには，広辞苑や新和英大辞典の記載に基づいて，複合語ごとに連濁形と非連濁形のどちらをとるのかが収録されている。データベース中では，連濁形には +，非連濁形には - が割り振られている。連濁の生起率を各後部要素に対して計算するために，+ を 1，− を 0 に変換することで数値化し，後部要素ごとの平均値を連濁生起率とした。なお，広辞苑と新和英大辞典の記載が一致しない場合は，0.5 とみなして連濁生起率の計算を行った。また，単一の辞書に連濁形と非連濁形の両方が記載されている場合も，0.5 とみなして連濁生起率の計算を行った。

2.3 データ分析

データの分析は，統計解析ソフトウェア R（バージョン 3.2.1，https://www.r-project.org/）により行った。まず，名詞，動詞由来名詞，動詞のそれぞれで，連濁生起率の分布に差があるのかどうかを 2 標本に対するコルモゴロフ・スミルノフ検定により調べた。さらに，連濁生起率の分布をいくつのサブグループに分類した場合に適合度が最大となるのかを，R の mclust パッケージ（バージョン 5.0.1，https://cran.r-project.org/web/packages/mclust/index.html）に含まれる Mclust 関数により検討した（Fraley and Raftery 2002）。mclust パッケージは，混合正規分布モデルに基づくクラスター分析を行うためのパッケージである。クラスター分析の適合度指標には，ベイズ情報量規準（Bayesian information criterion, BIC）を用いた。

（10）　$\mathrm{BIC} = -2\ln(L) + k\ln(n)$

L は尤度関数，k は独立変数の数，n は標本数

BIC は最適なモデルを統計的に選択する場合に一般的に使われる指標であり，BIC の値が小さ

いほど, モデルとの適合度が高いと解釈される[3]。クラスター分析では 1 個から 5 個の要素からなる混合正規分布を対象に，いくつの要素からなるモデルが最適かを検討した。本研究では，各正規分布の分散が等しい（つまり等分散の）混合正規分布を仮定したモデルと，分布ごとに分散が異なる（つまり不等分散の）混合正規分布を仮定したモデルを検討した。クラスター数が 1 個だけの時，等分散を仮定した場合と不等分散を仮定した場合で同じモデルとなるため，BIC は常に等しくなる。等分散の混合正規分布を仮定した場合は，それぞれの分布の平均値と全ての分布に共通の分散を推定する必要がある。これに対して，不等分散の混合正規分布を仮定した場合は，それぞれの分布に対して平均値と分散を独立に推定する必要があり，クラスター数が多い時には推定が収束しない可能性がある。

3. 結果

3.1 名詞，動詞由来名詞，動詞における連濁生起率の分布

後部要素が名詞，動詞由来名詞，動詞の場合に，連濁生起率の分布が有意に異なるかどうかを 2 標本に対するコルモゴロフ・スミルノフ検定により調べた結果，名詞と動詞由来名詞における連濁生起率の分布には有意差がなかった（$D = 0.12, p = 0.15$）。一方で，動詞における連濁生起率の分布は名詞や動詞由来名詞における分布と有意に異なっていた（名詞対動詞：$D = 0.72, p < 6.6^{-16}$，動詞由来名詞対動詞：$D = 0.78, p < 6.6^{-16}$）。これらの結果は，先行研究で名詞を対象に提案された，連濁生起率が 33% と 66% を境界として 3 つのサブグループに分かれるという仮説が，動詞に対しては成立しない可能性を示唆する。

3.2 名詞に対するクラスター分析

先行研究で検討された日本語複合名詞に対して，混合正規分布モデルに基づくクラスター分析を行った結果，分散が異なる 2 個の正規分布を用いた場合に最も適合度が高いモデルが得られた（BIC = 184.5）（次頁図 1）。不等分散の混合正規分布を仮定したモデルでは，混合正規分布中の各分布に対して，平均値と分散を推定する必要があり，クラスター数が 3 個以上の場合では推定が収束しなかった。そのため，クラスター数が 1 個と 2 個の場合の BIC のみを図に示した。この最適なモデルは，連濁生起率が 93.9 ± 0.38% と 46.6 ± 9.7% の 2 つの正規分布からなり，それぞれの正規分布には 264 個と 140 個の後部要素が分類された（次頁図 2）。Rosen（2001）や Irwin（2012）では, 33% と 66% という分類基準が提案されてきたが, 図 2 のヒストグラムから明らかなように, 連濁生起率が 90% 以上のサブグループと 50% 付近のサブグループに分類した場合に最適であり, この時のサブグループの分類基準は 90% 付近であることがわかった。一方で，Irwin（2012）が提案した通り，従来の *rendaku* waverer に分類される後部要素の頻度は全体の 10% 程度であった。

[3] 本研究で用いた mclust パッケージでは，$BIC = 2\ln(L) + k\ln(n)$ と定義されているため，BIC が最大のモデルが最適なモデルとなる。

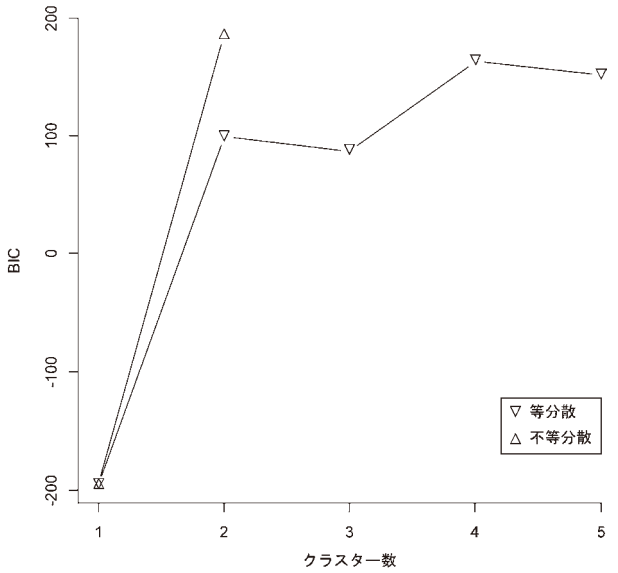

図 1　後部要素が名詞の場合に最適なクラスター数

縦軸は BIC，横軸はクラスター数を示す。クラスター数が 1 個の場合, 等分散と不等分散の BIC は常に等しい（以下の図 3，5，7 についても同様）。クラスター数が 2 個で不等分散の混合正規分布を仮定した場合に最も適合度の高いモデルが得られた。

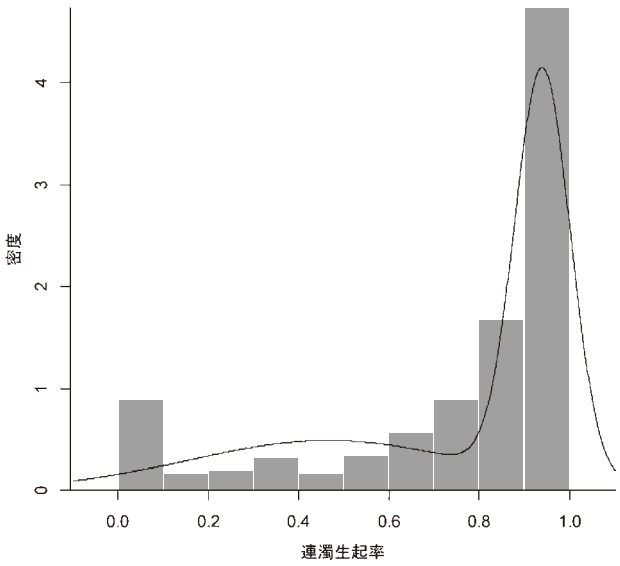

図 2　名詞における最適なモデルと連濁生起率のヒストグラム

縦軸は密度（相対度数），横軸は連濁生起率を示す。曲線は最適なモデルによる連濁生起率の予測を示す（以下の図 4，6，8 についても同様）。連濁生起率が 90% 以上のサブグループと 50% 付近のサブグループに分類された。

3.3 *Rendaku* immune を除外した名詞に対するクラスター分析

Rosen（2001）は，常に連濁を起こさない（つまり連濁生起率が 0% の）後部要素を “*rendaku* immune” と呼び，*rendaku* lover や *rendaku* hater と区別している。上記のクラスター分析では，*rendaku* immune な後部要素が含まれていたことが原因で，正しい結果が得られなかった可能性がある。そこで, *rendaku* immune である 17 の後部要素を除外したデータに対して, 同様のクラスター分析を再度実行した。その結果，*rendaku* immune を含んでいた図 1 および図 2 の結果と同様に，分散が異なる 2 個の正規分布を用いた場合に最も適合度が高いモデルが得られた（BIC = 256.2）（図 3）。この最適なモデルは，連濁生起率が 94.2 ± 0.35% と 53.7 ± 7.9% の 2 つの正規分布からなり，それぞれの正規分布には 259 個と 128 個の後部要素が分類された（次頁図 4）。この解析でもサブグループの分類基準は 80% を超えていることから，*rendaku* immune の影響で従来の研究と異なる結果が得られた可能性は低いと考えられる。

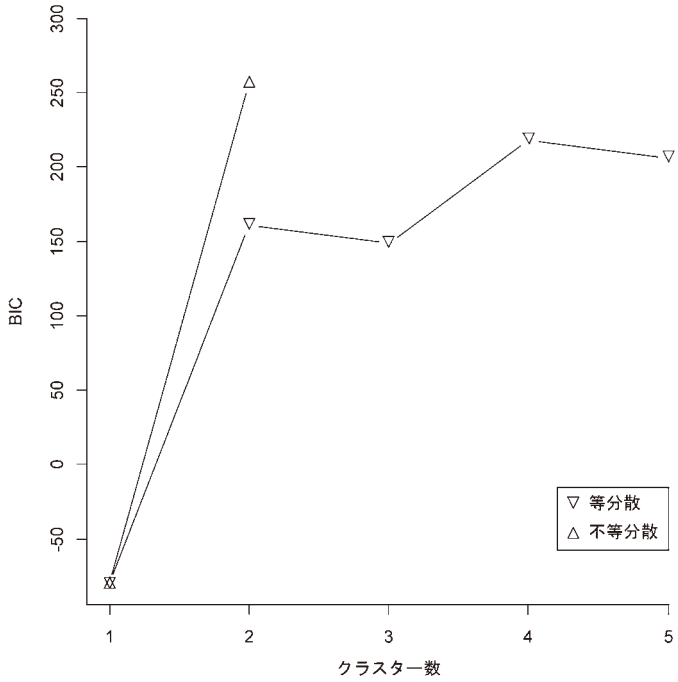

図 3　*Rendaku* immune を除外した名詞で最適なクラスター数
クラスター数が 2 個で不等分散の混合正規分布を仮定した場合に最も適合度の高いモデルが得られた。

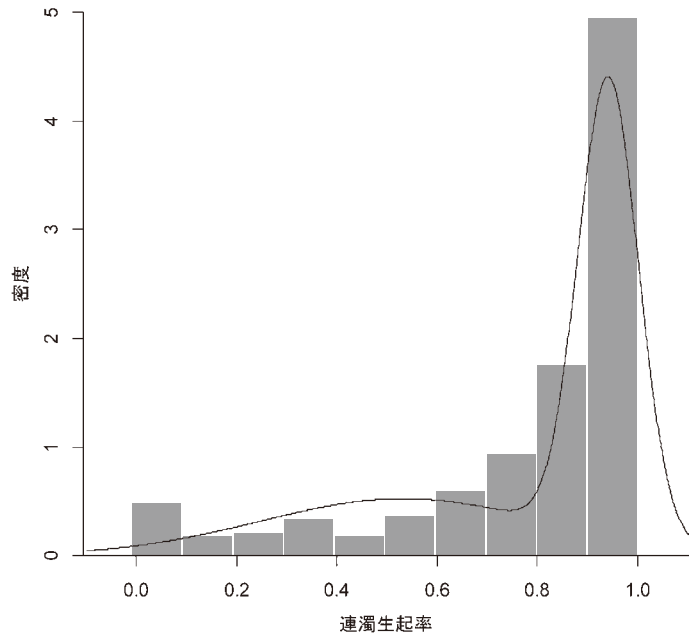

図 4　*Rendaku* immune を除外した名詞で最適なモデルと連濁生起率のヒストグラム
連濁生起率が 90% 以上のサブグループと 50% 付近のサブグループに分類された。

3.4 動詞由来名詞に対するクラスター分析

次に，動詞由来名詞に対するクラスター分析を行った結果，やはり分散が異なる 2 個の正規分布を用いた場合に最も適合度が高いモデルが得られた（BIC = 66.2）（図 5）。この最適なモデルでは，連濁生起率が 93.2 ± 0.36% と 56.4 ± 6.3% の 2 つの正規分布からなり，それぞれの正規分布には 98 個と 70 個の後部要素が分類された（図 6）。図 2 および図 6 のヒストグラムと，3.1 のコルモゴロフ・スミルノフ検定から明らかなように，動詞由来名詞の連濁生起率の分布は，名詞の分布と基本的に同様であった。Rosen（2001）や Irwin（2012）が対象としたのは派生を含まない名詞のみであったが，動詞由来名詞のように他の品詞から名詞に派生された後部要素を持つ複合語においても，派生を含まない名詞と同様の振る舞いを示すことが明らかとなった。以上の結果から，連濁の生起には派生前の品詞ではなく，派生後の品詞が重要であることが示唆された。

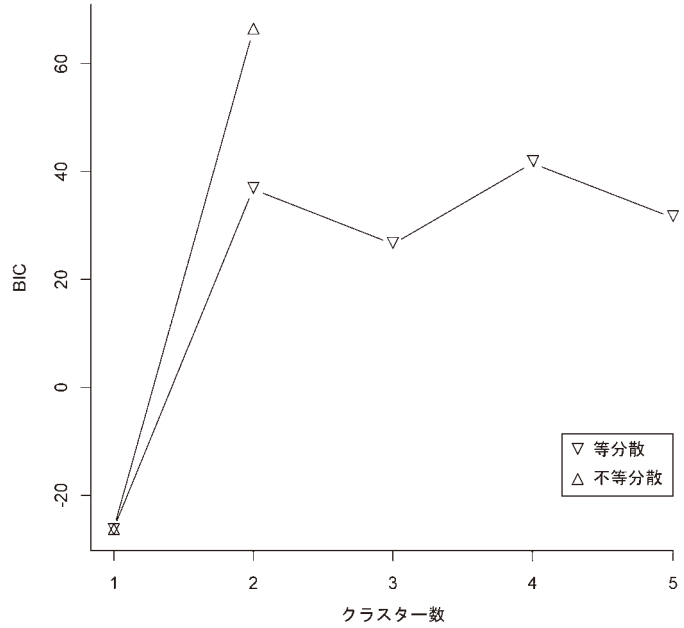

図 5　後部要素が動詞由来名詞の場合に最適なクラスター数
クラスター数が 2 個で不等分散の混合正規分布を仮定した場合に最も適合度の高いモデルが得られた。

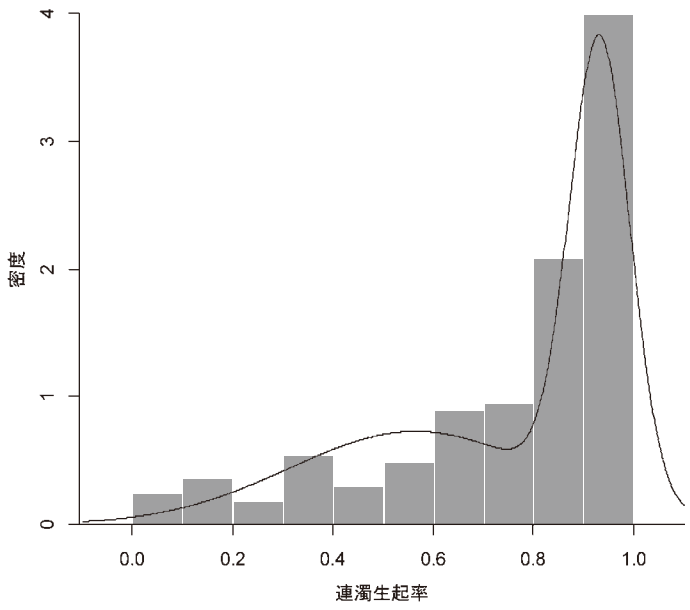

図 6　動詞由来名詞における最適なモデルと連濁生起率のヒストグラム
名詞の場合と同じく，連濁生起率が 90% 以上のサブグループと 50% 付近のサブグループに分類された。

3.5 動詞に対するクラスター分析

最後に，動詞に対するクラスター分析を行った結果，分散が等しい 2 個の正規分布を用いた場合に最も適合度が高いモデルが得られた（BIC = 40.3）（図 7）。この最適なモデルは，連濁生起率が 67.9 ± 0.97% と 6.3 ± 0.97% の 2 つの正規分布からなり，それぞれの正規分布には 8 個と 49 個の後部要素が分類された（図 8）。名詞や動詞由来名詞では，連濁生起率が 90% 以上の部分に分布のピークが存在したが，動詞では連濁生起率が 10% 未満の部分に分布のピークが存在した。サブグループの分類基準は 40% 付近であることがわかった。以上の結果から，動詞の連濁生起率の分布は，名詞や動詞由来名詞の分布と異なることが明らかとなった。先行研究においても，動詞では連濁が抑制されることが報告されており（伊東 2008），本研究の結果は先行研究の知見を再現したものと考えられる。

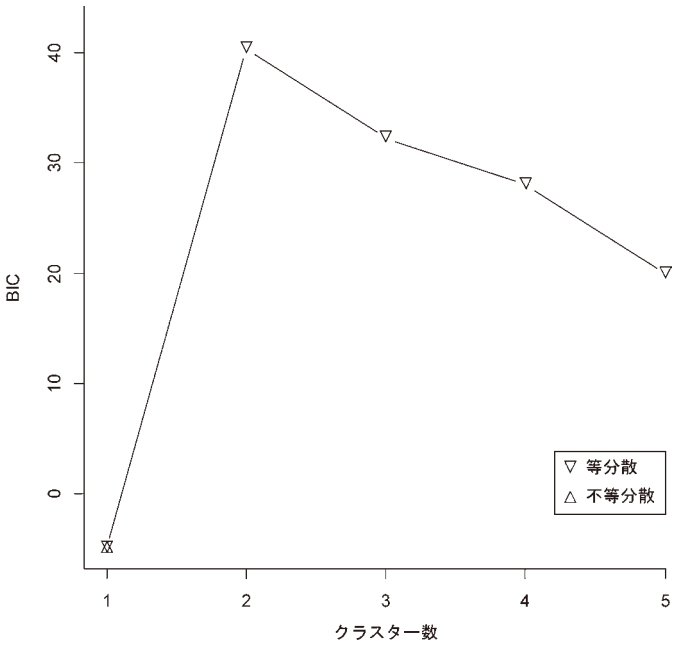

図 7　後部要素が動詞の場合に最適なクラスター数

不等分散の混合正規分布を仮定したモデルでは，クラスター数が 2 個以上の場合に分布の推定が収束しなかったため，クラスター数が 1 個の場合の BIC のみを図に示した。クラスター数が 2 個で等分散の混合正規分布を仮定した場合に最も適合度の高いモデルが得られた。

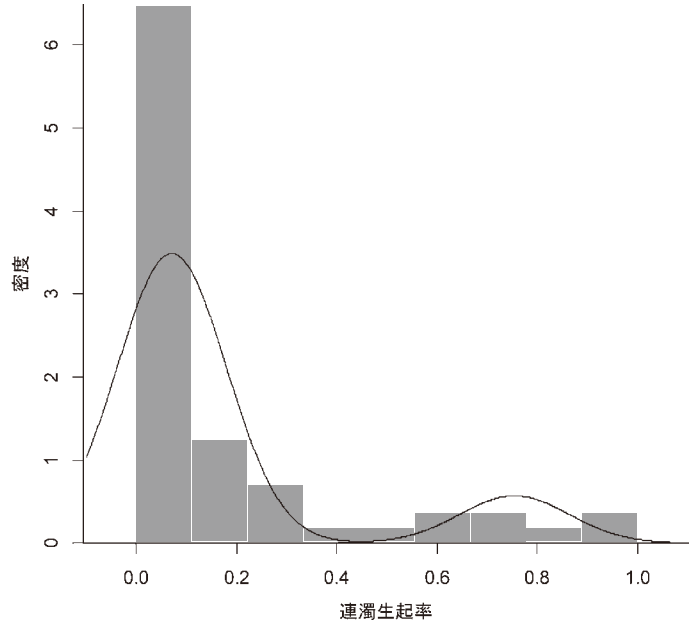

図 8　動詞における最適なモデルと連濁生起率のヒストグラム
名詞の場合と異なり，連濁生起率が 60% 以上のサブグループと 10% 未満のサブグループに分類された。

4. 考察と今後の展望

本研究では，後部要素が名詞，動詞由来名詞，動詞の日本語複合語に対して連濁生起率を調べ，いくつのサブグループに分類した場合に最適であるのかを，混合正規分布モデルに基づくクラスター分析により検討した。従来の研究では，連濁生起率が 33% 未満の後部要素を *rendaku* hater，33% から 66% の間の後部要素を *rendaku* waverer，66% を超える後部要素を *rendaku* lover と分類することが提案されてきた（Rosen 2001, Irwin 2012）。しかしこの分類の基準には恣意性があり，またサブグループの数をあらかじめ 2 個または 3 個と仮定している点にも問題があった。これらの問題に対処するため，クラスター分析によりデータの特性に基づいてサブグループの分類基準及びサブグループの数を決定した。

名詞及び動詞由来名詞に関しては，90% 以上の連濁生起率を持つグループと，50% 程度の連濁生起率を持つグループに分類した場合に最適なモデルが得られることが明らかとなった（図 2 と図 6）。この結果は，従来の *rendaku* hater と *rendaku* waverer は同じグループに属していること，また *rendaku* lover は従来よりもはるかに高い頻度で連濁を生じさせるグループであることを示すものである。後部要素が名詞の場合に，最適なモデルの分類基準やグループ数は，連濁が一切生じない *rendaku* immune を除外しても変化しなかったことから（図 3 と図 4），*rendaku* immune の影響で，従来の研究とは大きく異なる分類基準が得られた可能性は低いと考えられる。

動詞に関しては，先行研究から連濁が抑制されることが示唆されていた（伊東 2008）。コルモ

ゴロフ・スミルノフ検定の結果から，動詞の連濁生起率の分布は名詞や動詞由来名詞とは有意に異なることが示され，この結果は先行研究の知見を支持するものであった。さらに，クラスター分析の結果から，動詞の場合も 2 つのサブグループを仮定するモデルが最適であるが，その分類基準は名詞や動詞由来名詞と異なることも示された。動詞では連濁生起率が 10% 未満の部分に分布のピークが存在し，名詞や動詞由来名詞とは鏡像対称な分布であった（図 8）。以上の結果は，言語データの分類において，恣意性を排除して詳細な検討を行う場合にクラスター分析をはじめとする統計的手法が有効であることを示すものであった。太田（2015）や今回の結果から示唆されるように，今後は統計的手法に基づいて定量的にデータを分析する研究がさらに進展することが期待される。

本研究で用いた連濁データベースには，形容詞に由来する名詞や形容詞も含まれていたが，名詞や動詞に比べてこれらの品詞は収録数が少ないため分析の対象から除外した。出現頻度の低い品詞にクラスター分析を行うためには，現代日本語書き言葉均衡コーパス（BCCWJ）をはじめとする大規模コーパスを利用して研究を進める必要がある [4]。

また，混合正規分布モデルに基づくクラスター分析の適用範囲についても今後の検討を加える必要がある。混合正規分布モデルはさまざまなデータの分布に対応可能であり（McLachlan and Peel 2000），分布に対する事前知識がない場合にまず仮定するモデルとしては妥当であると考えられる。しかし連濁生起率のように 0% から 100% という有限の範囲を取る分布に対しては，ベータ分布などの他の分布を仮定した方が，適合度が上昇する可能性がある（Gupta and Nadarajah 2004）。今後はモデルで使用する分布についてもさらなる検討を加えたいと考えている。

参照文献

文化庁（編）（2014）『平成 25 年度「国語に関する世論調査」の結果の概要』. http://www.bunka.go.jp/kokugo_nihongo/yoronchousa/h25/pdf/h25_chosa_kekka.pdf (2015 年 6 月 5 日参照)

Fraley, Chris and Adrian E. Raftery (2002) Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association* 97: 611–631.

Gupta, Arjun K. and Saralees Nadarajah (2004) *Handbook of beta distribution and its applications* (Statistics: A Series of Textbooks and Monographs). New York: Marcel Dekker.

Irwin, Mark (2012) *Rendaku* haters and the logistic curve. Paper presented at 22nd Japanese/Korean Linguistics Conference.

Irwin, Mark (2014) *Rendaku* across duplicate moras. *NINJAL Research Papers* 7: 93–109.

Irwin, Mark and Mizuki Miyashita (2015) The Rendaku Database v2.5. http://www-h.yamagata-u.ac.jp/~irwin/site/Rendaku_Database.html (2015 年 3 月 5 日参照)

伊東美津（2008）「連濁について」『九州国際大学教養研究』15(2): 83–102.

Lyman, Benjamin S. (1894) The change from surd to sonant in Japanese compounds. *Oriental Studies*, 1–17. Oriental Club of Philadelphia.

McLachlan, Geoffrey and David Peel (2000) *Finite mixture models* (Wiley Series in Probability and Statistics). New York: Wiley & Sons, Inc.

中川芳雄（1966）「連濁・連清（仮称）の系譜」『国語国文』35(6): 302–314.

[4] BCCWJ は書き言葉コーパスであるが，発音形（出現形）と基本形（辞書形）が記載されているため，発音形と基本形で語頭に清濁の対立があれば連濁形，対立がなければ非連濁形としてコーパスを検索することで連濁のデータを得られる。

太田真理（2015）「音韻的・意味的要因が連濁に与える影響：連濁データベースとロジスティック回帰分析を利用した研究」『音韻研究』18: 85–92.
Otsu, Yukio (1980) Some aspects of *rendaku* in Japanese and related problems. In: Yukio Otsu and Anne Farmer (eds.) *Theoretical issues in Japanese linguistics* (MIT Working Papers in Linguistics 2), 207–227. Cambridge: MIT.
Rosen, Eric R. (2001) Phonological processes interacting with the lexicon: Variable and non-regular effects in Japanese phonology. Unpublished doctoral dissertation, University of British Columbia.
田中伸一（2009）『日常言語に潜む音法則の世界』（開拓社言語・文化選書 10）東京：開拓社.

Classification of Japanese Compounds Based on the Frequency of *Rendaku*: A Study Using the Rendaku Database

OHTA Satoshi[a]　OHTA Shinri[b]

[a]Yamaguchi University / Project Collaborator, NINJAL
[b]The University of Tokyo

Abstract
Rendaku is one of the most well-known phonological phenomena in Japanese, which voices the initial obstruent of the second element of a compound. Previous studies have proposed that Japanese compound words can be classified on the basis of the frequency of *rendaku* (*rendaku* rate). However, since these studies used arbitrary criteria to determine clusters, such as 33% and 66%, as well as arbitrary numbers of clusters, it is crucial to examine the plausibility of such criteria. In this study, we examined the optimal boundary criteria as well as the optimal number of clusters using a clustering analysis based on Gaussian mixture modeling and the Rendaku Database (Irwin and Miyashita 2015). The cluster analyses clarified that the two-cluster model was optimal for classifying both compound nouns and compound verbs. The boundary values of the *rendaku* rate for these clusters were approximately 90% and 40% for the compound nouns and compound verbs, respectively. These results were inconsistent with the findings of previous studies. Our findings demonstrate that model-based clustering analysis is an effective method of determining optimal classification of linguistic data.

Key words: *rendaku*, compound word, frequency, cluster analysis, Gaussian mixture model